

パラフレーズ技術を利用した 情報・知識利活用ソリューション

Information and Knowledge Utilization Solutions Using Paraphrasing Techniques

齋藤 佳美 倉田 早織 加納 敏行

■ SAITO Yoshimi ■ KURATA Saori ■ KANO Toshiyuki

情報・知識利活用ソリューションは、日常業務で作成される様々な業務文書において、文書の品質を高め、精度よく分類し整理して活用できるようにするためのシステムである。従来から、文書の検索や分類に単語を単位とした処理が行われているが、ことばの意味により近い高度な検索や分類へのニーズが高まっている。

東芝ソリューション(株)はこのようなニーズに応えるために、新たにパラフレーズ技術の開発に取り組んでいる。パラフレーズ技術とは、同じ意味を表す様々な表現を生成する技術であり、同じ意味の文を一つにまとめ上げるために利用できる技術である。従来の検索・分類技術をベースにパラフレーズ技術を活用することにより、文の意味を基にした処理を行い、高精度の検索や分類を実現する。

Information and knowledge utilization solutions manage a variety of business documents created and stored every day, not only to enhance the quality of documents but also to classify and arrange them with high accuracy. Searching and classification of documents have been conventionally processed in terms of the meaning of words. However, there is growing demand for more sophisticated technologies for semantic information that can support precise searching and classification of business documents according to the meanings of phrases.

In response to this situation, Toshiba Solutions Corporation has been engaged in the research and development of paraphrasing techniques that can generate and summarize various expressions representing the same meaning. We are now aiming to realize precise searching and classification of business documents according to the meanings of phrases by using paraphrasing techniques in addition to conventional technologies.

1 まえがき

東芝ソリューション(株)は、知識経営を支援するために情報・知識利活用技術の研究開発に取り組んでいる。

当社は、情報・知識利活用を「企業内外に蓄積されている多種多様な大量のデータから、有用な情報を迅速かつ的確に抽出し、業務の品質や効率の向上、リスクの低減、戦略の立案や意思決定に生かすこと」と定義している⁽¹⁾(図1)。このうち、必要な情報を収集する際に用いるのが文書検索技術、情報を共有し分析する際に用いるのが文書分類技術、文書品質を保持向上する際に用いるのが文書チェック技術である。

従来から、文書の検索や分類には、形態素解析によって文書を単語に分解し、単語を単位として文書を処理する手法が用いられている。しかし近年、文書量の爆発的な増加と文書の多様性の増大により、「誰・何が何をどうした」という文の意味を高精度に把握することの重要性が高まっている⁽²⁾。文の意味を高精度に把握するには、次のような課題がある。

- (1) 言語表現の多様性 同じ意味を持つ様々な表現が存在する。
- (2) 表現の省略や代名詞など、照応参照 複数文にまたがるような照応参照が存在する。

そこで当社は、(1)の課題を解決することを目的として、パラ

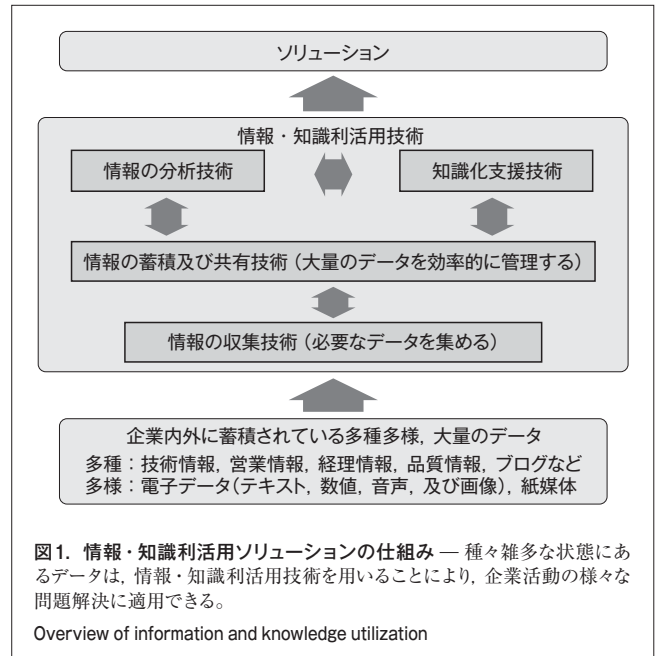


図1. 情報・知識利活用ソリューションの仕組み — 種々雑多な状態にあるデータは、情報・知識利活用技術を用いることにより、企業活動の様々な問題解決に適用できる。

Overview of information and knowledge utilization

フレーズ技術の研究開発に取り組んでいる。パラフレーズ技術とは、同じ意味を表す様々な文を生成する技術であり、同じ意味の文を一つにまとめ上げるために利用できる技術である。ここでは、当社が開発した、着目語を定めてパラフレーズ表現

を生成する着目語駆動型パラフレーズ技術の特長と、情報・知識利活用への適用方式について述べる。

2 着目語駆動型パラフレーズ技術の特長

2.1 パラフレーズ表現の例

パラフレーズとは、言い換え、換言とも呼ばれる言語現象で、その種類と例文を表1に示す。

実際には、複数種類のパラフレーズが複数箇所に生じるため、(7)のように表現の対応関係は複雑なものとなる。

2.2 従来のパラフレーズ技術とその課題

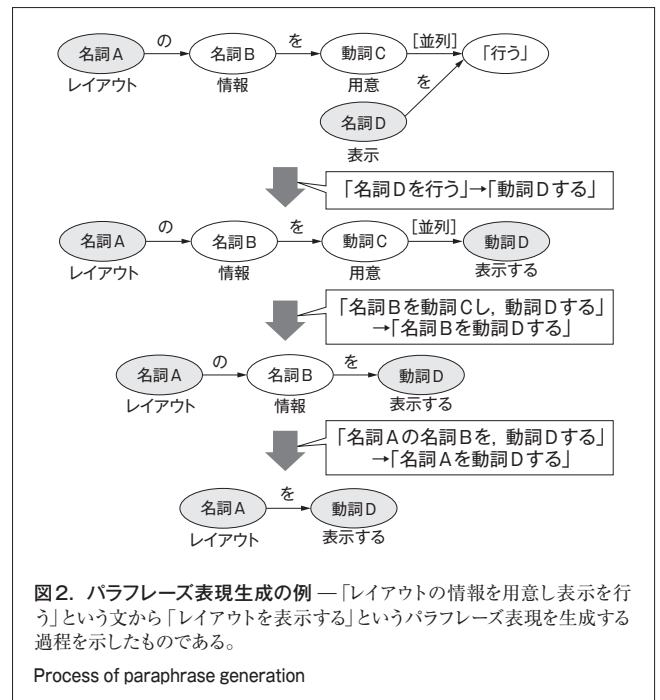
2.2.1 従来のパラフレーズ技術 従来、パラフレーズ技術には次のようなものがあった⁽³⁾。

- (1) パラフレーズ表現の生成技術 元となる表現からパラフレーズ表現を生成し、生成した表現を評価する。元となる表現の構文などを解析し、パターンマッチングルールによりパラフレーズ表現を生成する手法が多く用いられる。
- (2) パラフレーズ表現の発見技術 二つの表現の共通性を用いてパラフレーズ表現の候補を発見し、候補となる表現どうしを対応付ける。共通部分を多く含む表現を見つけ、二つの表現間をパターンマッチングルールにより対応付ける手法が多く用いられる。

2.2.2 従来技術の課題 前述のように従来は、パターンマッチングルールによりパラフレーズ表現を生成したり対応付けたりという手法が用いられている。このため、次のような課題がある。

- (1) 複雑なパラフレーズの生成や発見が難しい。複雑なパラフレーズに対応できるためには、複雑なパターンマッチングルールを用意する必要がある。

種類	表現例	
(1) 節内	プログラムの設計を進めた	プログラムの設計が進められた
	契約条件を記載した	記載した契約条件
	表紙の印刷を実行する	表紙を印刷する
(2) 文体及び機能表現	雨が降りそうだ	雨が降るだろう
(3) 内容語句	CO ₂ 排出量	温室効果ガス排出量
	日本放送協会	日本放送協会 (NHK)
(4) 節間	面積670 km ² の琵琶(びわ)湖は日本でいちばん大きい	琵琶(びわ)湖の面積は670 km ² だ。日本でいちばん大きい
(5) メトニミー(換喩(かんゆ))	シェイクスピアが書いた本を読む	シェイクスピアを読む
(6) 照応参照による	すべての文字列に対して同様の操作を行う	すべての文字列に対してタグ付けを行う
(7) 複雑なパラフレーズ	新製品の説明資料を作成し、3月3日の営業会議にてプレゼンテーションを実施した	昨日のミーティングで新商品のプレゼンを行った



- (2) 対象表現のどの部分がパラフレーズされるか特定できないため、多くのパラフレーズ表現が生成される。
- (3) パラフレーズ表現の生成や対応付けに用いられるパターンマッチングルールの記述に手間が掛かる。

2.3 着目語駆動型パラフレーズ技術の特長

当社は、2.2.1項で述べた技術を融合し、次のような着目語駆動型パラフレーズ技術を開発した⁽⁴⁾。

- (1) パラフレーズ表現を生成する対象文を決定し、その文の中でパラフレーズ表現の生成に際し着目する語(以下、着目語と呼ぶ)を定める。
- (2) パラフレーズ表現を生成するための生成規則を対象文に適用し、パラフレーズ表現を生成する。

この際、次の手順に従う。

- (a) 対象文に対し、生成規則を再帰的に適用する。
- (b) 対象文の着目語自体が削除されてしまわないような生成規則を適用する。

パラフレーズ表現の生成過程の例を図2に示す。

この方式の特長は、次のとおりである。

- (1) 単純な規則の組合せにより複雑なパラフレーズ表現を生成することができる。
- (2) 着目語を定めることにより、不要なパラフレーズ表現は生成されない。

3 パラフレーズ技術の情報・知識利活用への適用

3.1 文書検索

着目語駆動型パラフレーズ技術を適用した文書検索システム

ムを開発した。システムは次のような仕組みで動作する。

- (1) 入力された検索要求に基づきキーワード検索を行う。
- (2) キーワード検索結果から、別のパラフレーズ表現を生成する。この際、検索時にキーワードとして用いた単語を着目語とする。
- (3) 得られたパラフレーズ表現と検索要求とを比較し、類似度を評価する。
- (4) 一定の類似度を越えたパラフレーズ表現の元となった文だけを検索結果とする。

この文書検索システムの適用により、不要な文書が検索されなくなり、目的の文書を見つけやすくなる効果がある(図3)。

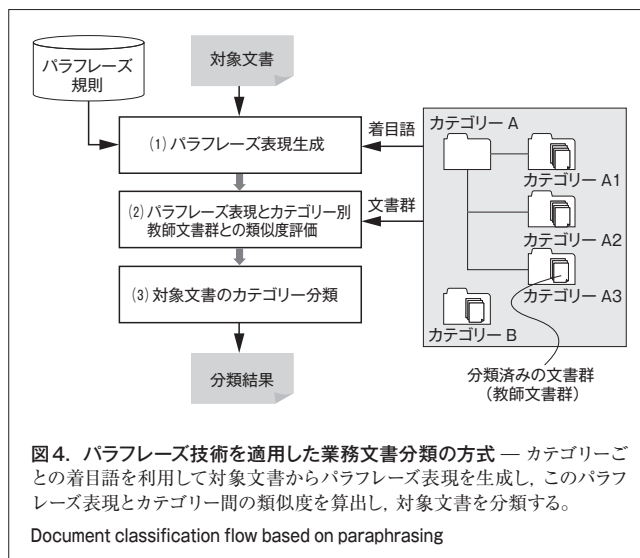
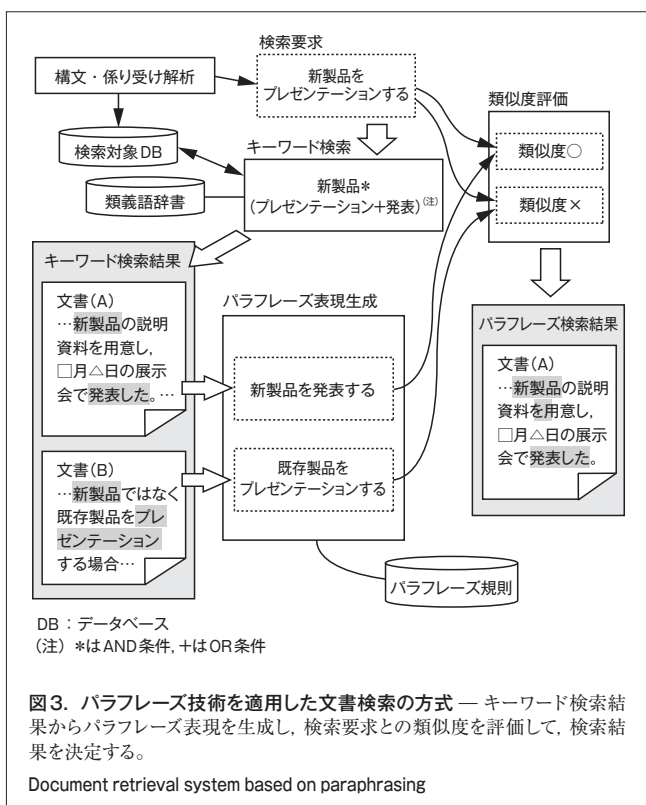
特許文書に対して行った評価実験では、検索件数がキーワード検索の約半分に絞り込まれ、かつ目的とする文の約90%が検索されるという結果が得られている。

3.2 文書分類

大量の業務文書を分析し知識にするためには、文書を細かく整理することが有効である。よって、単語の区別ではできない、詳細な文書分類にニーズがある。

従来の文書分類において、文書中の単語頻度などから、文書を分類構造(カテゴリ)に分類する手法がある。しかし、この手法は詳細な文書分類を実現するには限界があり、文書の意味を考慮したカテゴリ分類の実現が必要となる。

当社は、このようなカテゴリ分類を実現する一つのアプローチとして、パラフレーズ技術をベースとした業務文書分類



技術の開発に取り組んでいる。着目語駆動型パラフレーズ技術を適用して開発した業務文書分類の方式を図4に示す。この方式は、以下の三つの処理から構成される。

- (1) パラフレーズ表現生成 分類済みのカテゴリごとの着目語を利用して、対象文書中の表現に対し着目語駆動型パラフレーズ技術を適用し、パラフレーズ表現を生成する。
- (2) パラフレーズ表現とカテゴリ別教師文書群との類似度評価 生成されたパラフレーズ表現とカテゴリ別教師文書群との類似度を算出することより、対象文書とカテゴリの類似度を算出する。
- (3) 対象文書のカテゴリ分類 対象文書とカテゴリの類似度により対象文書を分類する。

詳細な文書分類が必要とされる文書として、特許文献、アンケート回答、不具合情報などがある。これらの文書は、同じ意味の様々な表現が含まれる特徴を持つ。この業務文書分類の方式を、単語を用いた特許文献の文書分類に加えて適用した評価では、分類精度が5%向上し、詳細な文書分類ができるようになった。

この業務文書分類の方式を用いて、アンケート回答から生成したパラフレーズ表現を表2に示す。表3には、表2のパラフレーズ表現を用いたアンケート回答の分類結果を示す。表2のパラフレーズ表現は、“表示”を中心としたパラフレーズ表現である。表3では、生成されたパラフレーズ表現により、三つのカテゴリ(A1, A2, A3)に対象文書群が分類されている。従来の文書分類では、これらの対象文書群は、“表示”と“行き先”の語を含むため、一つのカテゴリに分類される。

よって、着目語駆動型パラフレーズ技術の適用により、意味を考慮した、詳細な文書分類ができる。このような文書分類の実現により、有効な情報や知識が抽出でき、情報・知識利用の実現に大きな効果をもたらす。

表2. アンケート回答のパラフレーズ表現

Paraphrases for questionnaire answers

対象文書 No.	文書内容	生成されたパラフレーズ表現
1	行き先や経由地の表示を行ってほしい	行き先を表示する
2	行き先の文字を大きく表示してほしい	行き先を表示する
3	目的地の情報を表示する装置が欲しい	目的地を表示する
4	表示したい時刻表のボタンを押すと、現在時刻から終電まで表示してほしい	時刻表を表示する
5	時刻表が行き先案内板に表示されると便利	時刻表を表示する
6	行き先をタッチパネルで押したら、切符の金額が表示されたいのに	料金を表示する
7	行き先料金の表示が見やすい	料金を表示する
8	表示されている料金には、行き先までの特急料金が含まれてほしい	料金を表示する

表3. アンケート回答の分類結果

Categorization of questionnaire answers

分類カテゴリー	対象文書 No.	文書内容
A1	分類済み文書	列車名だけでなく行き先を表示した方がよい
	1	行き先や経由地の表示を行ってほしい
	2	行き先の文字を大きく表示してほしい
	3	目的地の情報を表示する装置が欲しい
A2	分類済み文書	時刻表を始発から表示して欲しい
	4	表示したい時刻表のボタンを押すと、現在時刻から終電まで表示してほしい
	5	時刻表が行き先案内板に表示されると便利
A3	分類済み文書	料金と所要時間を表示してほしい
	6	行き先をタッチパネルで押したら、切符の金額が表示されたいのに
	7	行き先料金の表示が見やすい
	8	表示されている料金には、行き先までの特急料金が含まれてほしい

3.3 知識化支援に向けた取組み

3.3.1 用例ベースのパラフレーズ生成技術 情報・知識活用ソリューションでは、エンドユーザー自身がカスタマイズできることを要望される場合が多い。そこで、用例ベースのパラフレーズ生成技術にも取り組んでいる。

用例ベースのパラフレーズ生成技術を利用すると、利用者は所望のパラフレーズの実例を用例集として設定することにより、用例を模倣したパラフレーズ表現を得ることができる。また、パラフレーズの種類ごとに用例集を用意し、使用する用例集を切り替えることにより、多様な種類のパラフレーズに対応できる。

3.3.2 業務文書チェック技術への適用 業務文書チェック技術は、業務文書中に潜む不適切な表現を機械的に洗い出し、文書作成者に注意や修正を促すことで、文書の品質向上に寄与できる技術である⁽⁵⁾。

業務文書チェック技術とパラフレーズ技術の融合により、不適切な表現に対し、意味内容を保ったまま適切な別のパラフレーズ表現を生成して、文書の修正候補を提示できるようになる。

これにより、文書作成効率のいっそうの向上が期待でき、これから総合的な文書作成支援技術として発展させていく予定である。

3.3.3 対話型文書分類技術への適用 対話型文書分類技術とは、ユーザーとシステムの協調作業を通じて分類の意図を明確にし、これに基づいて文書を分類する技術である⁽⁶⁾。

文書分類で作成したカテゴリーに対し、着目語駆動型パラフレーズ技術を適用することにより、詳細な分類や、識別性・可読性が高い文で表現されたカテゴリー名の生成を対話形式で実現していく。

4 あとがき

当社は、着目語駆動型パラフレーズ技術の開発と、この技術を文書検索、文書分類などに適用する技術の開発に取り組み、「誰・何が何をどうした」という文の意味に基づいて、有用な情報を迅速かつ的確に発見し加工するための処理を実現してきた。

今後は知識化支援に向けて、更に取組みを進めていく。

文献

- (1) 早川ルミ, ほか. 日本語解析技術を活用した業務支援ソリューション開発への取組み. 東芝レビュー. 64, 2, 2009, p.30-34.
- (2) 黒橋禎夫, ほか. “構造的言語処理による情報検索基盤技術の構築”. 文部科学省科学研究費補助金 特定領域研究「情報爆発時代に向けた新しいIT基盤技術の研究」平成18年度成果報告会. 東京, 2007-01, A01-00-05.
- (3) 乾 健太郎, ほか. 言い換え技術に関する研究動向. 言語処理学会誌. 11, 5, 2004, p.151-198.
- (4) 齋藤佳美, ほか. “言い換え処理技術の文書検索システムへの適用”. 第14回言語処理学会年次大会 発表論文集. 東京, 2008-03, 言語処理学会. 2008, p.794-796.
- (5) ZU, Guowei, et al. "The Supporting Technology of Business Document Proofreading based on Intercultural Differences". CEC'07 and EEE'07. Tokyo, 2007-07, IEEE. 2007, p.91-98.
- (6) 宮部泰成. ユーザーの意図を反映した対話型文書分類技術. 東芝レビュー. 64, 2, 2009, p.58-59.



齋藤 佳美 SAITO Yoshimi

東芝ソリューション(株) IT技術研究所 研究開発部 研究主務。自然言語処理分野の研究・開発に従事。情報処理学会, 言語処理学会会員。
Toshiba Solutions Corp.



倉田 早織 KURATA Saori, Ph.D.

東芝ソリューション(株) IT技術研究所 研究開発部, 理博。自然言語処理分野の研究・開発に従事。情報処理学会, 言語処理学会会員。
Toshiba Solutions Corp.



加納 敏行 KANO Toshiyuki

東芝ソリューション(株) IT技術研究所 研究開発部 研究主務。情報・知識活用技術の研究・開発に従事。日本OR学会, 言語処理学会会員。
Toshiba Solutions Corp.